PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-091305
(43)Date of publication of application : 04.04.1997

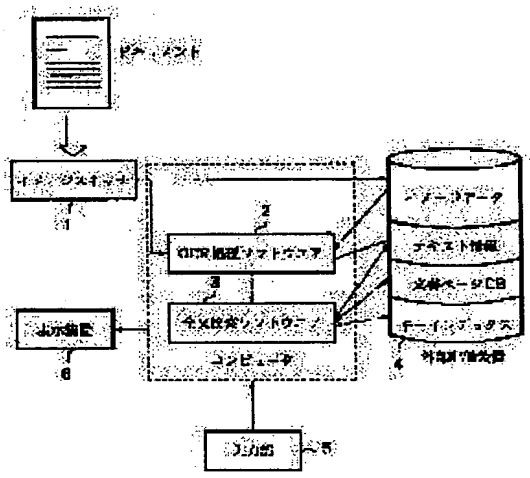
(51)Int.CI. G06F 17/30

(21)Application number : 07-249499
(22)Date of filing : 27.09.1995

(71)Applicant : CANON INC
(72)Inventor : SHIIYAMA HIROTAKA

(54) METHOD AND DEVICE FOR INFORMATION PROCESSING

(57)Abstract:
PROBLEM TO BE SOLVED: To effectively narrow down information by storing information, representing the positions of word keys and character keys in key indexes by as a small amount of information as possible.
SOLUTION: Document data are divided into plural page areas, which are managed by a document pages DB. Each page area is further divided into plural small areas, and key indexes wherein the page areas where information representing the respective keys in the document data and the small areas is registered are generated and stored in an external storage device 4. When retrieval is performed, a key index is retrieved on the basis of a key obtained by decomposing a retrieval word specified through an input part 5 to extract the page area where all the keys of the retrieval word are present in the same small area in the same page area. The part of the document data corresponding to the extracted page area is obtained and the final retrieval of the retrieval word is performed to obtain a retrieval result.



LEGAL STATUS

[Date of request for examination] 26.09.2002

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平9－91305

(43)公開日　平成９年(1997)４月４日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30			Ｇ０６Ｆ	15/413	３１０Ｂ
				15/40	３７０Ａ
				15/403	３１０Ｃ

審査請求　未請求　請求項の数８　ＯＬ　（全９頁）

(21)出願番号　特願平7－249499

(22)出願日　平成７年(1995)９月27日

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目30番２号

(72)発明者　椎山　弘隆
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

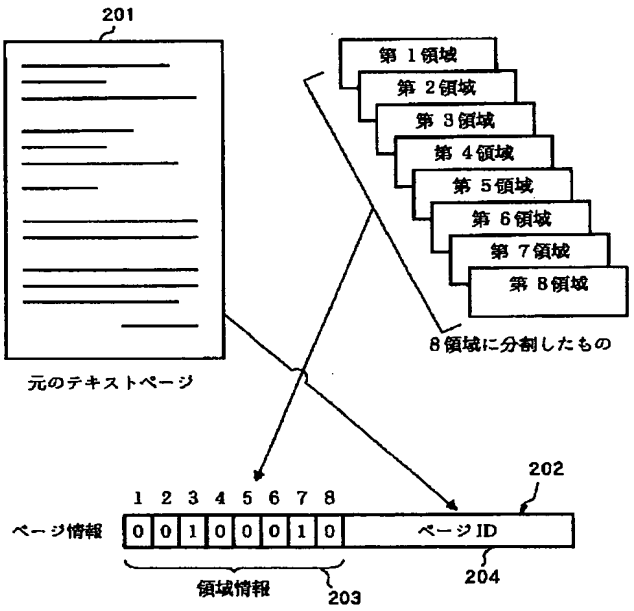
(74)代理人　弁理士　大塚　康徳　(外１名)

(54)【発明の名称】　情報処理方法及び装置

(57)【要約】

【課題】単語キーや文字キーの位置を示す情報をより少ない情報量でキーインデックスに記憶し、効果的な絞り込みを行う。

【解決手段】文書データは複数のページ領域に分割して文書ページＤＢによって管理される。複数のページ領域の各々は更に複数の小領域に分割され、文書データ中の各キーについて、各々のキーが存在するページ領域と小領域とを示す情報を登録したキーインデックスが生成されて外部記憶装置４に格納される。検索時においては、入力部５より指定された検索語を分解して得られたキーにより前記キーインデックスを検索し、同じページ領域中の同じ小領域に該検索語の全てのキーが存在するページ領域を抽出する。そして、抽出されたページ領域に該当する文書データの部分を獲得して前記検索語の最終的な検索を行い、検索結果を得る。



1

【特許請求の範囲】

【請求項１】　文書データを複数の領域に分割する分割手段と、
前記文書データより得られる各キーに対して、各々が存在する領域を示す情報を登録したキーインデックスを生成する生成手段と、
指定された検索語を分解して得られたキーによって前記キーインデックスを検索し、該検索語の全てのキーが同じ領域に存在する領域を抽出する抽出手段と、前記抽出手段で抽出された領域に対して前記検索語の検索を行い、検索結果を得る検索手段とを備えることを特徴とする情報処理装置。

【請求項２】　文書データを第１の領域単位で複数のページ領域に分割して管理する管理手段と、
前記複数のページ領域の各々について第２の領域単位で更に複数の小領域に分割する分割手段と、
前記文書データ中の各キーについて、各々のキーが存在するページ領域と小領域とを示す情報を登録したキーインデックスを生成する生成手段と、
指定された検索語を分解して得られたキーにより前記キーインデックスを検索し、同じページ領域中の同じ小領域に該検索語の全てのキーが存在するページ領域を抽出する抽出手段と、
前記文書データの前記抽出手段で抽出されたページ領域に該当する部分を獲得して前記検索語の検索を行い、検索結果を得る検索手段とを備えることを特徴とする情報処理装置。

【請求項３】　前記生成手段において、前記ページ領域を示す情報はページ番号であり、前記小領域を示す情報は対応するビットのオン・オフで示され、
前記抽出手段において各キーが同じ小領域に存在するか否かは前記小領域を示す情報同士の論理積をとることで判断することを特徴とする請求項２に記載の情報処理装置。

【請求項４】　前記分割手段によって得られる小領域は、少なくとも同一ページ内で互いに重複する部分を有することを特徴とする請求項２に記載の情報処理装置。

【請求項５】　前記ページ領域において、当該領域中の文字数が所定量に満たない場合は、当該ページ中の複数の小領域を１つの小領域とみなすことを特徴とする請求項２に記載の情報処理装置。

【請求項６】　前記検索語の指定とともに、各ページ領域に共通の検索位置として所望の小領域を指定する指定手段を更に備え、
前記抽出手段は、前記検索語の全てのキーが存在する小領域として前記指定手段で指定された小領域を含むページ領域を抽出することを特徴とする請求項２に記載の情報処理装置。

【請求項７】　文書データを複数の領域に分割する分割工程と、

2

前記文書データより得られる各キーに対して、各々が存在する領域を示す情報を登録したキーインデックスを生成する生成工程と、
指定された検索語を分解して得られたキーによって前記キーインデックスを検索し、該検索語の全てのキーが同じ領域に存在する領域を抽出する抽出工程と、
前記抽出工程で抽出された領域に対して前記検索語の検索を行い、検索結果を得る検索工程とを備えることを特徴とする情報処理方法。

【請求項８】　文書データを第１の領域単位で複数のページ領域に分割して管理する管理工程と、
前記複数のページ領域の各々について第２の領域単位で更に複数の小領域に分割する分割工程と、
前記文書データ中の各キーについて、各々のキーが存在するページ領域と小領域とを示す情報を登録したキーインデックスを生成する生成工程と、
指定された検索語を分解して得られたキーにより前記キーインデックスを検索し、同じページ領域中の同じ小領域に該検索語の全てのキーが存在するページ領域を抽出する抽出工程と、
前記文書データの前記抽出工程で抽出されたページ領域に該当する部分を獲得して前記検索語の検索を行い、検索結果を得る検索工程とを備えることを特徴とする情報処理方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は文書データ（テキストデータを含む）から所望のテキストデータを検索する情報処理方法及び装置に関する。

【０００２】

【従来の技術】従来より、文書データの全体について検索を行う全文検索システムがある。この全文検索システムにおいては、単純にテキストデータ全体をなめるような処理では検索速度が遅くなるため、これを解決するための手段としてキーインデックスを作成している。キーとは、例えばテキストデータから抽出した単語、簡単なものでは１文字・２文字を単純に切り出したものであり、キーインデックスとはその切り出したキーがどのテキストファイルに存在するかを記憶したデータベースの一種である。

【０００３】

【発明が解決しようとする課題】しかしながら、検索語が存在する文章を検索する際、単語や１文字・２文字キーがテキストデータのどの位置に存在するか不明な場合、検索語と一致したものを絞り込むことは非常に困難である。

【０００４】例えば、「新聞紙」という言葉を検索する場合、１文字インデックスで、「新」、「聞」、「紙」の３文字が存在するテキストデータを検出しても実際は、「新たに…聞いたこと…紙に書いておきましょう」

という風に個々のキーがばらばらに離れていることがあり得る。このため、実際にはテキストデータを全部なめるまでは検索結果として確定できない。もちろん解決策として、１文字キーの位置を示すアドレス情報をインデックスに持たせる方法も存在するが、インデックスファイルの容量が巨大になるため実用上、採用は不可能である。

【０００５】この例に見られる様に、如何に位置情報の欠如を克服し、絞り込み最終確定のための全文なめ処理を少なくするかが、技術的に重要なポイントである。

【０００６】同じことが単語キーの場合にも言え、検索語が文章で入力された場合にも同様の問題が発生する。

【０００７】本発明は上記の問題に鑑みてなされたものであり、単語キーや文字キーの位置を示す情報をより少ない情報量でキーインデックスに記憶し、効果的な絞り込みを行うことが可能な情報処理方法及び装置を提供することを目的とする。

【０００８】

【課題を解決するための手段】上記の目的を達成するための本発明の情報処理装置は、文書データを複数の領域に分割する分割手段と、前記文書データより得られる各キーに対して、各々が存在する領域を示す情報を登録したキーインデックスを生成する生成手段と、指定された検索語を分解して得られたキーによって前記キーインデックスを検索し、該検索語の全てのキーが同じ領域に存在する領域を抽出する抽出手段と、前記抽出手段で抽出された領域に対して前記検索語の検索を行い、検索結果を得る検索手段とを備える。

【０００９】また、上記の目的を達成する本発明の他の構成による情報処理装置は、文書データを第１の領域単位で複数のページ領域に分割して管理する管理手段と、前記複数のページ領域の各々について第２の領域単位で更に複数の小領域に分割する分割手段と、前記文書データ中の各キーについて、各々のキーが存在するページ領域と小領域とを示す情報を登録したキーインデックスを生成する生成手段と、指定された検索語を分解して得られたキーにより前記キーインデックスを検索し、同じページ領域中の同じ小領域に該検索語の全てのキーが存在するページ領域を抽出する抽出手段と、前記文書データの前記抽出手段で抽出されたページ領域に該当する部分を獲得して前記検索語の検索を行い、検索結果を得る検索手段とを備える。

【００１０】また、好ましくは、前記生成手段において、前記ページ領域を示す情報はページ番号であり、前記小領域を示す情報は対応するビットのオン・オフで示され、前記抽出手段において各キーが同じ小領域に存在するか否かは前記小領域を示す情報同士の論理積をとることで判断する。小領域中に検索語の各キーが存在するか否かを、小領域を示す情報同士の論理積で判断できるので、キーの存在位置のより細かい分析を容易かつ高速に行えるからである。

【００１１】また、好ましくは、前記分割手段によって得られる小領域は、少なくとも同一ページ内で互いに重複する部分を有する。文書データを小領域に分割することによって連続して出現しているキーが分離されてしまう可能性が有るが、これを防止することができるからである。

【００１２】また、好ましくは、前記ページ領域において、当該領域中の文字数が所定量に満たない場合は、当該ページ中の複数の小領域を１つの小領域とみなす。例えばイメージやグラフなどの挿入により文字数が少ないページ領域では、これを小領域に分離すると連続したキーを分離してしまうなど、却って弊害を生じやすいが、これを防止できる。

【００１３】また、好ましくは、前記検索語の指定とともに、各ページ領域に共通の検索位置として所望の小領域を指定する指定手段を更に備え、前記抽出手段は、前記検索語の全てのキーが存在する小領域として前記指定手段で指定された小領域を含むページ領域を抽出する。各ページに共通の検索位置を指定して検索を実行することが容易に実現できる。

【００１４】上記の構成によれば、文書データが複数の領域に分割され、文書データより得られる各キーに対して、各々が存在する領域を示す情報を登録したキーインデックスが生成される。そして、指定された検索語を分解して得られたキーによってこのキーインデックスを検索し、該検索語の全てのキーが同じ領域に存在する領域を抽出する。抽出された領域に対して前記検索語の検索を行い、検索結果を得る。このように、検索語による最終的な検索に先立って、検索位置の絞り込みが行われるので、検索処理の速度が向上する。

【００１５】また、上記の他の構成によれば、文書データは第１の領域単位で複数のページ領域に分割して管理される。そして、複数のページ領域の各々を第２の領域単位で更に複数の小領域に分割し、文書データ中の各キーについて、各々のキーが存在するページ領域と小領域とを示す情報を登録したキーインデックスを生成する。検索時においては、指定された検索語を分解して得られたキーにより前記キーインデックスを検索し、同じページ領域中の同じ小領域に該検索語の全てのキーが存在するページ領域を抽出する。そして、抽出されたページ領域に該当する文書データの部分を獲得して前記検索語の最終的な検索を行い、検索結果を得る。このように検索語による最終的な検索に先立って検索位置の絞り込みが行われる。特に、小領域中における検索語キーの存在を調べるので、効果的な絞り込みができる。更に、文書データをページ領域と小領域の２段階の領域で分割するので、段階的な絞り込みが可能となり、絞り込みの処理効率が向上する。

【００１６】

5

【発明の実施の形態】以下に添付の図面を参照して本発明の好適な実施形態を説明する。

【００１７】図１は本実施形態の情報処理装置のテキスト検索に係る制御構成を表すブロック図である。同図において、１はスキャナであり、文書を光学的にスキャンしてイメージ情報を得る。２はＯＣＲ処理ソフトウエアであり、イメージスキャナ１もしくは外部記憶装置４より得られたイメージデータについて文字認識処理を行い、テキスト情報を得る。３は全文検索ソフトウエアであり、件テキスト情報からキーインデックスを作成し、検索処理を行う。４は外部記憶装置であり、イメージ情報とテキスト情報および検索用ファイル等を記憶する。５は入力部であり、検索語、検索条件を入力するためのキーボードやマウス等から構成される。６は表示装置であり、検索語・検索条件を入力するための表示およびイメージデータを表示したりする。

【００１８】本実施形態の情報処理装置は、蓄積・登録処理と検索処理を実行する。

【００１９】蓄積・登録処理は、（１）文書ページＤＢに新規登録文章の登録およびページ情報の設定、（２）イメージスキャナ１から得られたイメージ情報を外部記憶装置４に記憶する作業、（３）イメージ情報をＯＣＲ処理ソフトウェア２でテキスト情報化した後に外部記憶装置４に記憶する作業、および（４）テキスト情報から本件アルゴリズム（図３に示す）に従ってキーインデックス作成処理を行う全文検索ソフトウェア３で作成したキーインデックスを外部記憶装置４に記憶する作業等からなっている。

【００２０】又、検索処理は、（１）キーボード５から入力された検索語を全文検索ソフトウェア３が受け取り、登録時と同じアルゴリズムでキー分解した後、各キーに対応するページ情報をキーインデックスから読み込み、本件アルゴリズム（図４に示す）でページＩＤによる第一次絞り込み、領域情報のＡＮＤ処理による第二次絞り込みを経た後、最終的にテキストデータをなめて検索結果を確定する作業、（２）文書ページＤＢから文書アドレス情報を取り出す作業、および（３）表示装置６に該当イメージデータを表示する作業からなる。

【００２１】では、次に本件アルゴリズムによる登録・検索処理の具体的な例を挙げて説明を行う。

【００２２】まず、蓄積・登録時において、本実施形態においては、テキスト・文書データは複数のページからなるものとし、複数のページファイルに分割して各々にユニークなページＩＤを付けてテキスト・文書データを格納する。そして、各ページともとのテキスト・文書データとの対応を文書ページＤＢによって管理する。なお、ページという概念が存在しないテキストデータでは、文字数や行数によって仮想的にページ分けする。この文書ページＤＢはページＩＤによるもとのテキスト・文書データと個々のページとの対応だけではなく、テキスト・文書データの属性情報、例えば文書名や日付け、所有者等を記憶し、文書属性による検索にも用いることが可能である。

6

【００２３】上記登録時において、単語キーや文字キーがページのどの位置に存在するかを表す１バイト～数バイトの領域情報を採用する。これは、１ページを複数領域に分割し、そのキーが存在する領域に対応するビットを立てたものであり、ページＩＤに領域情報を付加したもの（以降ページ情報と呼ぶ）をキーインデックスのそのキーに対応するレコードに記憶する。

【００２４】図２は本実施形態におけるページ情報を説明する図である。元のテキストページ２０１は第１領域から第８領域の８つの領域に分割される。ここで、図示のように各領域は互いにある程度重なりあうものとし、領域境界による不都合を解消する。２０２はページ情報であり、領域情報２０３とページＩＤ２０４とを含む。領域情報２０３は１ページ内の領域分割数に対応したビットを有し、後述の検索処理で検索文字が見つかった領域のビットが１にセットされる。図の例では、第３領域と第７領域に検索文字が存在することを示す。このようなページ情報が各キーに付与される。

【００２５】また、各領域の大きさは各々ページの文字数または行数と領域数により決定する。もし１ページの文字数または行数が少ない場合には、領域情報は全てのビットが立ったもの（本例では０ｘｆｆ（１６進数のｆｆ）がセットされる）とする。以下に登録処理を図３のフローチャートを参照して更に説明する。

【００２６】図３は本実施形態の登録処理の手順を表すフローチャートである。なお本処理はテキスト・文書データにおいてページ単位の分割が終了した後に、１ページごとに起動されるものとする。従って、図３のフローチャートでは１ページ分の登録処理が示されている。

【００２７】まず、ステップＳ１１において１ページ中の文字数を取得する。そして、１ページ中の文字数と分割数（本例では８こ）等に基づいて分割領域の大きさを設定する。ステップＳ１３で未読み込みの領域が存在すればステップＳ１４へ進み、設定された分割領域の大きさ分だけデータの読み込みを行う。そして、ステップＳ１５においてキー分解処理を行う。キー分解処理とは、読み込んだデータを１文字もしくは２文字、或は単語等のキーに分解し、各キーに対してページ情報を付与するものである。なお、１つの領域中に複数のキーが存在する場合は、２つ目以降のキーについてはページ情報の付与を行わない。即ち、１つの領域においては、１つのキーに対して１つのページ情報が割り当てられるようにする。

【００２８】以上の処理を当該ページの全ての領域について実行すると未読み込みの領域が存在しなくなるので処理はステップＳ１３からステップＳ１６へ進む。

【００２９】ステップＳ１６では、当該ページにおいて

7

複数の領域に存在するキーを一つのページ情報にまとめる。例えば、図２に示したように、第３領域と第７領域にキーが存在した場合は、領域情報２０３の対応するビットをセットする。続いて、ステップＳ１７において領域情報を上位、ページＩＤを下位に格納した各キーのページ情報をキーインデックスに登録する。

【００３０】以上のような処理を全ページについて実行することにより、当該文書データに対するキーインデックスが形成される。

【００３１】次に、上記のキーインデックスを用いた本実施形態の検索処理について説明する。

【００３２】図４は本実施形態のキーインデックスの構成例と検索手順の概要を説明する図である。同図では、上位の１バイトを領域情報とし、下位の３バイトをページを指定するためのページＩＤ番号とする、計４バイトのページ情報を要素に持つキーインデックス中の登録内容が示されている。なお、ページＩＤとして３バイトを割り当てているが、これは、中規模ファイリングシステムではページにユニークな番号を振っても３バイトあれば足りるからである。ページ情報のバイト数構成は上記に限らないことは言うまでもない。

【００３３】領域情報が１バイトの場合、領域は８領域となり、そのキーが存在する領域に対応したビットが１にセットされる。もし１ページの行数または文字数がある値より少なければ領域情報を０ｘｆｆとして処理することにより、領域分割の弊害を防ぐ。

【００３４】次に、キーインデックスを用いて検索処理が実行される。検索処理では、まずインデックスレコードの情報中のページＩＤを見て、全てのキーに対するインデックスレコードで同じページＩＤを持つ、即ち１つのページ中に検索後を分解したキー全てが揃っているページ情報を個々のキーに対して取り出す。これを第一次絞り込みと呼ぶ。

【００３５】次に、取り出されたページ情報の領域情報の部分を見て同じビットが立っている、即ち同じ領域に検索語を分解したキー全てが揃っているページ情報を取り出し、有効なページ情報として保存する。これを第二次絞り込みと呼ぶ。

【００３６】分解したキーが存在するだけ、上記の第一次・第二次絞り込み処理およびこの結果と前回の有効なページ情報と共通なものを新たなページ情報として保存する。このような処理を繰り返した最終結果が最終的な第二次絞り込み結果となり、このページＩＤから文書ページＤＢに照会し、対応するテキストデータを取り出し、全文をなめて確認した結果が最終検索確定結果となる。

【００３７】さて、図４の例を見ると、「製」「品」という１文字キー２個に対するインデックスが示されている。また、このキーインデックスは上記の登録処理によって生成されたものである。例えば、ページＩＤ番号０

8

ｘ１２３４５６のページをキー分解した結果、これら「製」「品」の２文字が含まれていたことがわかり、更に、上位１バイトにはその文字が当該ページ中の８領域のどこに存在しているかを示す領域情報が格納される。

【００３８】上記の如きキーインデックスを用いて、例えば「製品」という言葉で検索処理を実行した場合、まず「製」「品」各々のキーに対して、これらのキーを持つページ情報の配列（インデックスレコード）をキーインデックスから得る。

【００３９】そして、この２つのキーのページ情報（４バイト）配列の中身を見て、両方に存在するページＩＤを抽出することで、第一次絞り込みを行う。これは互いのページ情報配列のページＩＤ部分（下位３バイト）のみを総当たりで論理積演算した結果に相当する。

【００４０】次に、上記第一次絞り込みで得た各々の文字キーに対するページ情報で、同じページＩＤを持つものの領域情報（上位１バイト）同士でビット毎の論理積演算を行う。この結果、１個でも同じ位置のビットが立っていたもの、即ち同じ領域に文字キーが存在しているものを得ることで第二次絞り込みを行う。

【００４１】これらの第一次・第二次絞り込みで、「製」「品」の２キーが同じページの同じ領域に存在するページのみに絞り込むことが出来、最終確定のための全文なめの対象が大幅に絞り込める。

【００４２】この様に、少ない情報量ではあるが、キーの存在する領域情報をキーインデックスに持たせることにより、非常に効率的に最終確定のための全文なめの作業量を少なくすることが可能となる。

【００４３】図５は本実施形態の検索処理の手順を表すフローチャートである。まず、ステップＳ２１において入力部５を用いて検索語を指定する。ステップＳ２２では指定された検索語をキーに分解する。そして、ステップＳ２３において、１個目のキーに対応するインデックスレコードを有効なページ情報としてキーインデックスから読み込む。

【００４４】ステップＳ２４において未処理のキーが存在するならば処理はステップＳ２５へ進み、その未処理のキーの一つに対応するインデックスレコードをキーインデックスから読み込む。ステップＳ２６では、ステップＳ２３で得た有効なページ情報とステップＳ２５で読み込んだインデックスレコードの各ページ情報とを比較し、同じページＩＤを有するページ情報を保存する。即ち、ステップＳ２６では第一次絞り込みが行われる。

【００４５】次にステップＳ２７では、ステップＳ２７で保存されたページ情報の領域情報部分同士のビット毎の論理積をとり、結果が０でないページ情報を有効なページ情報として保存する。即ちステップＳ２７では第二次絞り込みが行われる。以上のような処理を指定された検索語を分解して得られた全てのキーについて行うと、処理はステップＳ２４からステップＳ２８へ進む。

9

【００４６】ステップＳ２８では最終的に残ったページ情報のページＩＤのテキストデータを文書ページＤＢを参照して外部記憶装置４からロードし、全文をなめて確認し、検索語の存在したページ情報のみを保存する。そして、ステップＳ２９において、最終検索結果として最終的に残ったページ情報を出力する。

【００４７】なお、領域情報を生かしたものとして、検索時に検索語がページのどの位置にあるかを指定して検索する単語位置曖昧指定検索を行うこともできる。この場合、例えばステップＳ２１において検索語とともに検索位置を指定する。そして、ステップＳ２７において第二次絞り込みを行う際に、指定領域に対応したビットを立てた指定領域データと、ステップＳ２６の第一次絞り込みで得たページ情報中の領域情報との論理積を取ることで実現できる。

【００４８】また、上記図５のフローチャートでは、１キー毎に第一次絞り込み、第二次絞り込みを行うがこれに限られるものではない。例えば、全てのキーに対応するページ情報をロードした後にまとめて第一次絞り込み、第二次絞り込みを行うようにしてもよいことは言うまでもない。

【００４９】以上のように、本実施形態では、インデックスレコードの情報中のページＩＤを見て、全てのキーに対するインデックスレコードで同じページＩＤを持つ、即ち１つのページ中に検索後を分解したキー全てが揃っているページ情報を個々のキーに対して取り出す第一次絞り込みと、取り出されたページ情報の領域情報の部分を見て同じビットが立っている、即ち同じ領域に検索語を分解したキー全てが揃っているページ情報を取り出し、有効なページ情報として保存する第二次絞り込みとが実行される。そして、検索語を分解して得られたキーの全てについて上記の第一次・第二次絞り込み処理をくり返し、このくり返しの過程で有効なページ情報を絞り込み、得られたページ情報のページＩＤから文書ページＤＢに照会し、対応するテキストデータを取り出し、全文をなめて確認した結果を最終検索確定結果とする。

【００５０】従って本実施形態によれば、テキストデータを得る手段により得られた大量のテキストデータを蓄積している記録媒体から効率的且つ高速にテキストデータを検索することが可能となる。

【００５１】なお、上記実施形態では文書・テキストデータを管理するシステムへの適用を説明したが、これ以外にも、文書画像からＯＣＲにより得たテキストデータに対する全文検索システムによる画像検索、更には文字データを含まない画像データに対しても付加した説明テキストデータを対応付けておくことにより検索可能な画

10

像ファイリングシステムにも応用可能である。

【００５２】もちろん複数検索語とその論理演算指定、シソーラス（類義語）展開した検索語の処理も、本発明のアルゴリズムによる各々の検索結果を演算すれば可能である。

【００５３】以上のように本実施形態によれば、単語キーや文字キーの位置情報をそのまま記憶するのではなく、単語キーや文字キーがページのどの位置に存在するかを表す数バイトの領域情報を採用することにより、検索語を分解した全てのキーが同じページの同じ領域に存在するページのみに絞り込むことが出来、非常に効果的に最終確定のための全文なめの対象を絞り込むことが可能となり、結果として大幅に検索速度を向上できる。

【００５４】更に領域情報を生かしたものとして、検索時に検索語がページのどの位置にあるかを指定して検索する単語位置曖昧指定検索も可能となる。

【００５５】尚、本発明は、複数の機器から構成されるシステムに適用しても、１つの機器から成る装置に適用しても良い。また、本発明はシステム或は装置にプログラムを供給することによって達成される場合にも適用できることはいうまでもない。

【００５６】

【発明の効果】以上のように本発明によれば、単語キーや文字キーの位置を示す情報をより少ない情報量でキーインデックスに記憶し、効果的な絞り込みを行うことが可能となり、検索処理速度を向上できる。

【００５７】

【図面の簡単な説明】

【図１】本実施形態の情報処理装置のテキスト検索に係る制御構成を表すブロック図である。

【図２】本実施形態におけるページ情報を説明する図である。

【図３】本実施形態の登録処理の手順を表すフローチャートである。

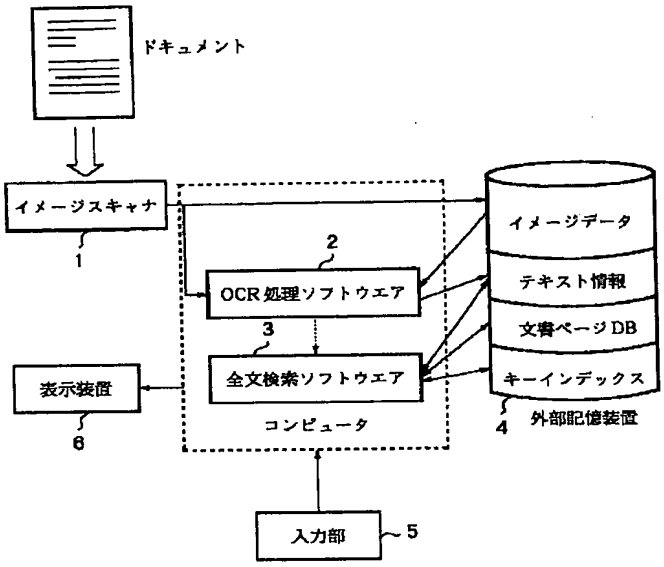
【図４】本実施形態のキーインデックスの構成例と検索手順の概要を説明する図である。

【図５】本実施形態の検索処理の手順を表すフローチャートである。

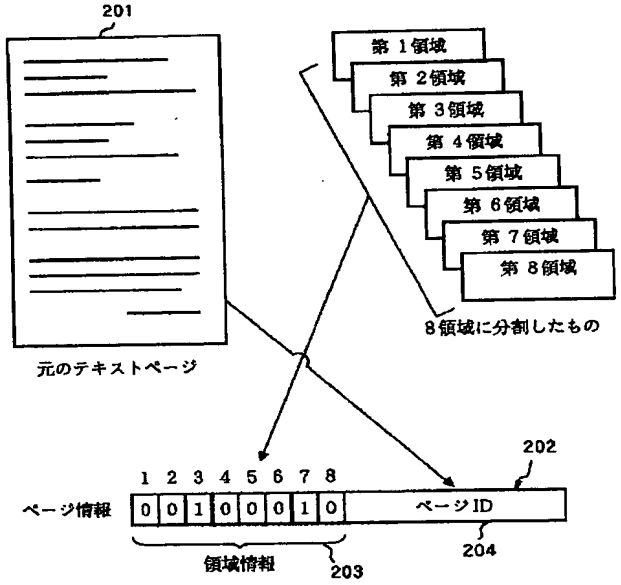
【符号の説明】

１　イメージスキャナ
２　ＯＣＲ処理ソフトウェア
３　全文検索ソフトウェア
４　外部記憶装置
５　キーボード
６　表示装置

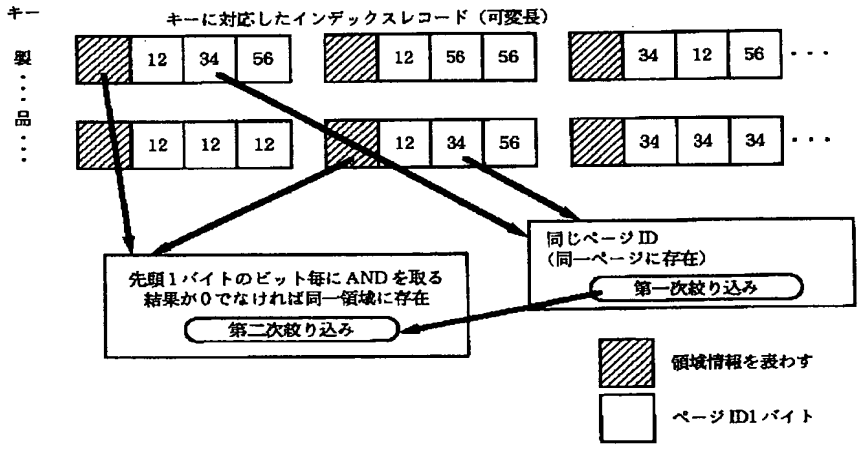
【図１】


ドキュメント
イメージスキャナ
1
2
OCR処理ソフトウエア
3
全文検索ソフトウエア
表示装置
6
コンピュータ
入力部
5
イメージデータ
テキスト情報
文書ページDB
キーインデックス
4
外部記憶装置

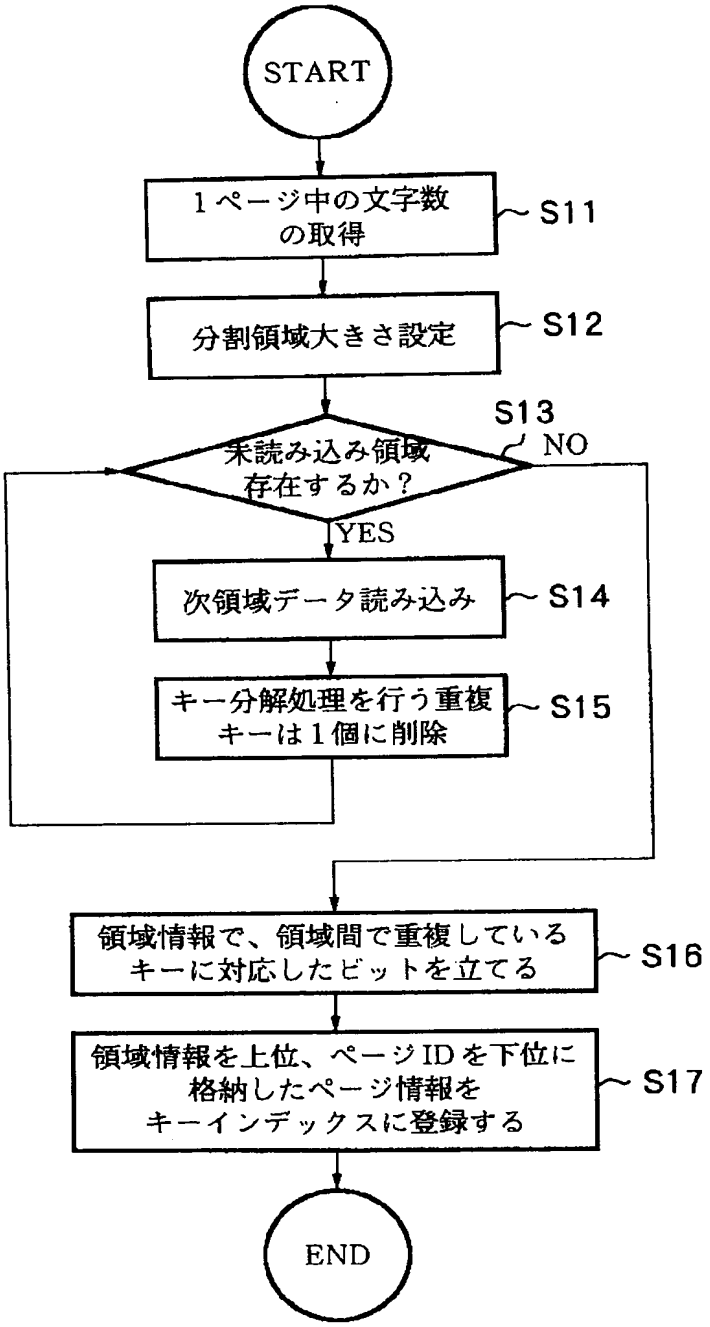
【図２】


201
元のテキストページ
第１領域
第２領域
第３領域
第４領域
第５領域
第６領域
第７領域
第８領域
８領域に分割したもの
202
1 2 3 4 5 6 7 8
ページ情報
0 0 1 0 0 0 1 0
ページID
領域情報
203
204

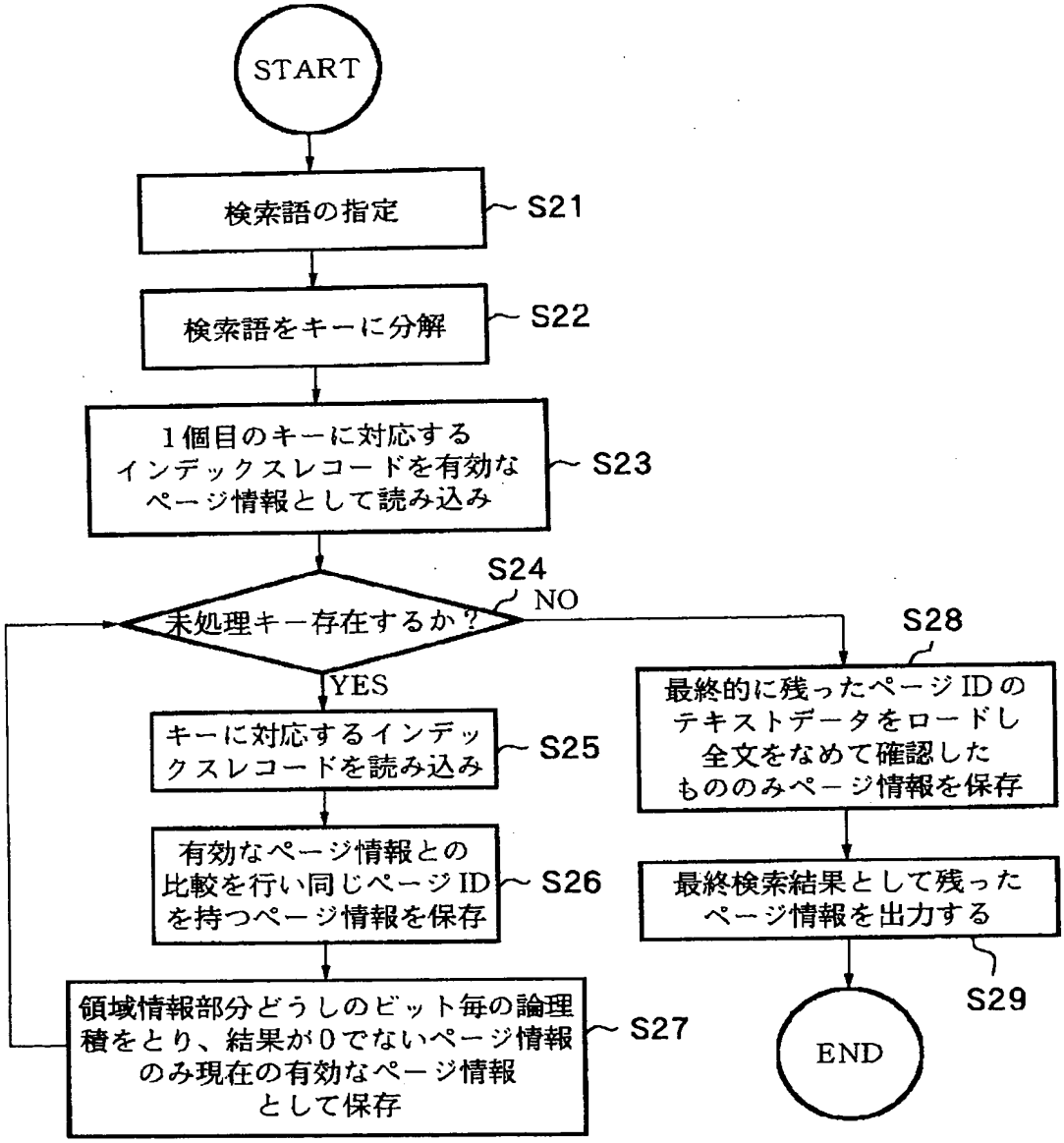
【図４】


キー
製品
キーに対応したインデックスレコード（可変長）
12 34 56
12 56 56
34 12 56
12 12 12
12 34 56
34 34 34
先頭１バイトのビット毎にANDを取る
結果が０でなければ同一領域に存在
第二次絞り込み
同じページID
（同一ページに存在）
第一次絞り込み
領域情報を表わす
ページID1バイト

【図３】


START
１ページ中の文字数の取得
S11
分割領域大きさ設定
S12
S13
未読み込み領域存在するか？
NO
YES
次領域データ読み込み
S14
キー分解処理を行う重複キーは１個に削除
S15
領域情報で、領域間で重複しているキーに対応したビットを立てる
S16
領域情報を上位、ページIDを下位に格納したページ情報をキーインデックスに登録する
S17
END

【図５】


START
検索語の指定
S21
検索語をキーに分解
S22
１個目のキーに対応する
インデックスレコードを有効な
ページ情報として読み込み
S23
S24
未処理キー存在するか？
NO
YES
S28
最終的に残ったページIDの
テキストデータをロードし
全文をなめて確認した
もののみページ情報を保存
キーに対応するインデッ
クスレコードを読み込み
S25
有効なページ情報との
比較を行い同じページID
を持つページ情報を保存
S26
最終検索結果として残った
ページ情報を出力する
S29
領域情報部分どうしのビット毎の論理
積をとり、結果が０でないページ情報
のみ現在の有効なページ情報
として保存
S27
END

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☑ FADED TEXT OR DRAWING
- ☑ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☑ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.